



COMP450 Applied Computing Project

A DOCUMENT PROCESSING SOLUTION FOR THE LOGISTIC INDUSTRY

Group members:

s1155180@live.hkmu.edu.hk (Eraka Yip)

s1265005@live.hkmu.edu.hk (Timothy Chan)

s1284810@live.hkmu.edu.hk (Andy Lau)

Project Aim

The aim of this project is to develop a document processing solution for the logistics industry that leverages Optical Character Recognition (OCR) and machine learning methodologies to reduce manual, repetitive input tasks among collaborating parties and enhance data security.

Project Objectives

The objectives of our project include the following points:

1. Preprocessing of logistics company documents in pdf file to image format using OpenCV module. Image processing and transformation may be necessary to enhance the accuracy of our trained model.
2. Building a machine learning model (e.g. CNN) using the collected dataset; Classification of documents by company name or project name using our trained model and/ or open-source object detection model. Save the documents in appropriate shared folders to prevent unauthorized access.
3. Use of OCR engine and machine learning model to convert structured/unstructured data from pdf documents to digital format and extract the booking number, the key identifier of a document for tagging.
4. Based on access privileges, allow only authorized users to get access to booking data.

Project Background

1. In the logistics industry, import and export are the two major but challenging activities. A complete operation involves collaboration of different parties, with each of them having different systems and data policy. Lots of manual work is needed in the workflow due to lack of system integration, which decreases efficiency and data integrity.
2. Meanwhile, hundreds of documents are generated every day. These documents have similar key information but may vary in the format, including company logo and the general layout. Companies have to classify documents by company or by project name to smoothen their operations.
3. As a platform provider of the import / export system (Party C), we have been working hard to achieve better collaboration among parties in the industry. The following table illustrates the roles and relationships among parties.

	Party A	Party B	Party C	Party D
Role	Vessel and Service Provider for import/export	Provide Storage Service for import/export	Our Company. Platform provider for import/export system	Request Booking
Description & relationships	Various global companies	- Party A is their customer - Need to serve Party D.	- Party B is their customer - Need to serve both Party A and Party D.	End user of import/export platform.

Table 1: Roles and relationships among parties in a typical import / export operation.

Apparently, we can get Party D data from part A directly to match with booking data in the platform and the problem should be solved. Unfortunately, Party A cannot disclose any Party D information because of business agreements. In addition, party A involves various companies and each company has its own data policy and system. It is time-consuming and costly to deal with case by case. So we need to solve this without any party's help.

With great effort, our company successfully establishes a common platform for all parties and helps consolidate data from them (see Figure 1). End users (Party D) can use data on our platform to complete a full operation. However, despite working, there are some critical issues regarding the current platform:

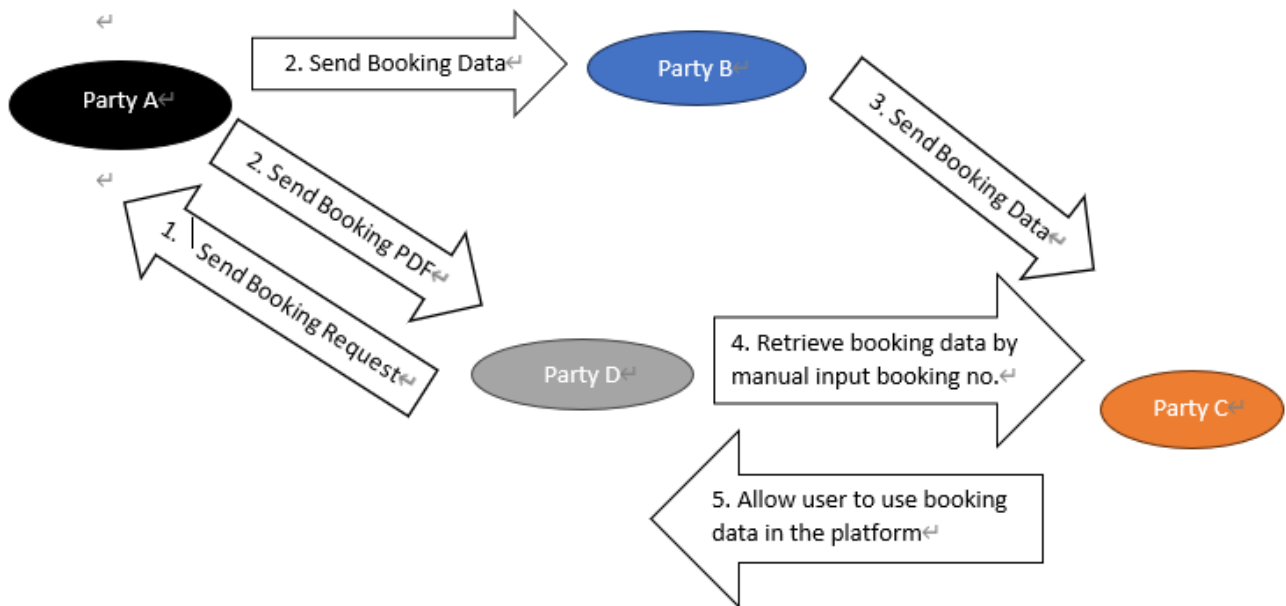


Figure 1: Import / Export workflow among different parties

Problem facing:

- a. Data security issue - Party D can obtain unauthorized data by inputting a wrong booking number intentionally or an unintentionally mistake created by manual input.
 - b. Too many human mistakes slow down operation efficiency - Irreversible impact may occur if a company in party D uses the incorrect data to do further actions. Also, after a company retrieves a booking, the booking will be locked. Therefore, incorrect manual input may cause inconvenience to other companies.
 - c. Various document formats increase human mistakes.
4. Given the current issues, we decide to develop:
- a. A classification method based on machine learning (e.g. CNN) for document identification among companies to handle documents in various formats generated from different companies.
 - b. A document processing model that helps us identify the booking number of the document, which is the key identifier of a document using ML techniques such as CNN and OCR. It helps reduce manual input error.

Project value proposition

1. Enhance document security.
2. Reduce manual data input and error.
3. Improve efficiency.
4. Enhance customer satisfaction.